DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 33, No. 3, 2017, pp. 625–657, http://dx.doi.org/10.1515/JOS-2017-0030

Dynamic Question Ordering in Online Surveys

Kirstin Early[1], *Jennifer Mankoff*[2], *and Stephen E. Fienberg*[3]

Online surveys have the potential to support adaptive questions, where later questions depend on earlier responses. Past work has taken a rule-based approach, uniformly across all respondents. We envision a richer interpretation of adaptive questions, which we call Dynamic Question Ordering (DQO), where question order is personalized. Such an approach could increase engagement, and therefore response rate, as well as imputation quality. We present a DQO framework to improve survey completion and imputation. In the general survey-taking setting, we want to maximize survey completion, and so we focus on ordering questions to engage the respondent and collect hopefully all information, or at least the information that most characterizes the respondent, for accurate imputations. In another scenario, our goal is to provide a personalized prediction. Since it is possible to give reasonable predictions with only a subset of questions, we are not concerned with motivating users to answer all questions. Instead, we want to order questions to get information that reduces prediction uncertainty, while not being too burdensome. We illustrate this framework with two case studies, for the prediction and survey-taking settings. We also discuss DQO for national surveys and consider connections between our statistics-based question-ordering approach and cognitive survey methodology.

Key words: Adaptive survey design; cognitive aspects of survey methodology; cost-based dynamic question ordering; questionnaire design.

1. Introduction

Survey response rates have been falling for decades, leading to results that do not necessarily represent the full population of interest (Porter 2004). Online surveys tend to have much lower response rates than traditional mail-out/mail-back and telephone surveys (Shih and Fan 2008). Unlike these traditional-styled surveys, online surveys can easily support adaptive question ordering, where the order of later questions depends on responses to earlier questions. Past work in adaptive questions for online surveys has taken a rule-based, question-specific approach where a certain response to a certain question leads to a new set of questions, uniformly across all respondents (e.g., (Pitkow and Recker 1995; Bouamrane et al. 2008)). We envision a richer interpretation of adaptive question ordering, where question order is dynamic and personalized to the individual, depending on their previous answers. Such a dynamic question-ordering approach has the potential of

[1] Machine Learning Department, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, U.S.A. Email: kearly@cs.cmu.edu
[2] Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, U.S.A. Email: jmankoff@cs.cmu.edu
[3] Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, U.S.A. Email: fienberg@stat.cmu.edu

increasing engagement, and therefore response rates, as well as imputation quality for missing values.

In addition to using surveys to gain insights about general populations, we can use survey results to provide individual respondents with useful information. We consider a respondent's answering a sequence of questions to receive a personalized estimate as a type of survey too. An example of such a survey is a carbon calculator, where a user provides information about their home infrastructure and energy consumption to get an estimate of their carbon footprint (Pandey et al. 2011). Here the respondent receives useful information from the survey, has a personal incentive to complete the survey, and likely self-selects into the survey. It is possible that the user does not need to answer all questions to get an accurate estimate. For some users, certain features will be more relevant than for other users. We can lower the cost (to users) of providing answers by ordering the questions so that the most informative questions for a particular user are asked first.

We present a general framework for dynamically ordering the questions that make up a survey questionnaire, based on previous responses, to engage respondents and improve survey completion and imputation of unknown items. Our work considers two scenarios for data collection from survey-takers. In the first, we want to maximize survey completion (and the quality of necessary imputations) and so we focus on ordering questions to engage the respondent and collect hopefully all the information we seek, or at least the information that most characterizes the respondent so imputed values will be accurate. For example, there may be a latent structure to a respondent's answers, and this underlying structure could guide question selection. In the second scenario, our goal is to give the respondent a personalized prediction, based on information they provide. Since it is possible to give a reasonable prediction with only a subset of questions, we are not concerned with motivating the user to answer all questions. Instead, we want to order questions so that the user provides information that most reduces the uncertainty of our prediction, while not being too burdensome to answer.

Any statistics-based approach to dynamic question ordering of the sort we consider here would seem to run counter to traditional arguments that questionnaires should have a fixed structure for all respondents and when the same quantities, for example, unemployment or poverty, are measured by surveys over time. Just over thirty years ago, the Cognitive Aspects of Survey Methodology (CASM) movement, for example, see (Jabine et al. 1984; Sudman et al. 1996; Tanur 1992), made the argument that this traditional approach to survey design shackled respondents and often prevented them from providing the very answers that the survey methodologists sought for their questions, for example, see (Suchman and Jordan 1990; Tanur 1992). We believe our approach reopens the door to the arguments raised by that movement, but in a very different manner, and somehow survey statisticians will ultimately need to blend the lessons from the CASM movement with the needs for cost-driven dynamic ordering. One important problem is that order effects (Sudman et al. 1996) can influence how people interpret and answer certain questions, leading to inconsistent data across respondents who received questions in different orders. However, not all questions are susceptible to order effects, and it is possible to constrain dynamically ordered questionnaires to respect relative orderings among sensitive questions (see Section 4).

The remaining sections of the article are as follows: first, in Section 2, we review related work in survey burden, respondent engagement, and question ordering from a variety of

fields. Then, in Section 3, we illustrate the promise of DQO with two case studies: the first collects data to provide personalized energy estimates to prospective tenants, drawing on data from the Residential Energy Consumption Survey (U.S. Energy Information Administration 2009), and the second is a traditional survey collection approach, with the Survey of Income and Program Participation (U.S. Bureau of the Census 2014b). In Section 4 we formalize and generalize the DQO framework, beyond the particular applications in the previous section. In Section 5 we set forth a broader view of the forms dynamic question ordering can take in other national surveys and suggest how they might benefit from a dynamic question-ordering approach. Finally, in Sections 6 and 7 we summarize our contribution and note avenues for future work in this area.

2. Related Work

In this section we first review the literature on survey burden, respondent engagement, and data quality. Then we cover work from several fields on adaptive question ordering, for personalizing the order in which information is acquired to achieve higher quality data at lower costs.

2.1. Survey Burden, Respondent Engagement, and Data Quality

Respondent burden in surveys is multifaceted, with multiple factors influencing what is ultimately a subjective measure for the respondent. Some of these factors are survey length (e.g., number of questions or estimated time), respondent effort, respondent stress, or frequency of interviews (Bradburn 1978). A respondent's perception of the survey task has a significant direct effect on self-reports of survey burden (Fricker et al. 2014). Recent work has demonstrated that telling respondents that they have been screened into a longer or shorter survey, with a longer or shorter expected time commitment, can influence their perception of survey burden. Those who were told they were screened into a longer survey reported more burden than those who also received the longer survey but were not informed of their selection (Yu et al. 2015). Reported survey length is one factor that influences a respondent's decision to begin a survey: fewer people start surveys that are announced to take more time to complete (e.g., (Walston et al. 2006; Marcus et al. 2007; Galesic and Bosnjak 2009)). However, this phenomenon is not universal (e.g., (Cook et al. 2000)). Furthermore, response quality tends to decrease as the survey progresses (e.g., (Galesic 2006; Barge and Gehlbach 2012)).

As survey response rates have been dropping (e.g., (Porter 2004; Shih and Fan 2008)), researchers have been looking at how to motivate respondents to fill out these surveys, to avoid having survey results not represent the full population. Commercial surveys often pay respondents, but compensation does not necessarily ensure thoughtful responses – participants still exhibit satisficing behavior in paid surveys (e.g., (Barge and Gehlbach 2012; Kapelner and Chandler 2010)). Incentivizing respondents with something dependent on the quality of their answers, like a personalized prediction or calculation, can motivate them to provide correct data. For example, Angelovska and Mavrikiou (2013) design an online questionnaire that gives the respondent feedback on their level of procrastination, based on their responses. Their experiments find that the questionnaire that promises feedback has lower breakoff rates than the standard questionnaire in which

the respondent does not receive personalized feedback. Marcus et al. (2007) find that offering personalized feedback increased response rates for low-salience surveys but had no significant effect for high-salience surveys.

Another promising venue for increasing respondent motivation and lowering breakoff is by using *paradata* collected as the respondent answers an online survey (e.g., time spent on page, mouse clicks (Kaczmirek 2008)) to model user engagement (Couper et al. 2010). Survey breakoff is influenced by respondent factors (e.g., demographics), survey design (e.g., topic), and page and question characteristics (e.g., question type) (Peytchev 2009). Respondent interest and burden are negatively and positively, respectively, associated with breakoff, and respondent and survey factors influence these components too (Galesic 2006). Furthermore, lower-quality answers (e.g., skipped items) often precede breakoff (Galesic 2006). Survey action can be taken to increase user engagement and response rates.

2.2. *Adaptive Question Ordering*

There is a rich literature focusing on adaptively ordering questions to improve outcomes while minimizing respondent burden, across multiple fields. Examples include adaptive design in survey methodology, adaptive treatment design in medical statistics, adaptive testing in educational research, and test-time feature selection in machine learning.

2.2.1. Adaptive Survey Design

Adaptive survey design (ASD) attempts to improve survey quality (in terms of achieving a higher response rate or lower error) by giving respondents custom survey designs, rather than the same one (Schouten et al. 2013). Usually ASD tries to minimize nonresponse, and designs involve factors like number of follow-ups, which can be costly. The general technique is to maximize survey quality while keeping costs below a budget.

Often in ASD, changes in survey design happen between *phases* of the survey, where the exact same survey protocol (e.g., sampling frame, survey mode, measurement conditions) is in place within a phase and results from that phase inform changes to the protocol for the next phase. Groves and Heeringa (2006) introduce an approach they call responsive survey design, which uses indicators of the cost and error of design features to make decisions about how to change the survey design in future phases and then combines data from all phases into a final estimator. They also introduce the concept of *phase capacity*—once a stable estimate has been reached in a design phase, it is unlikely that expending more effort in that phase will result in a better estimate. Their definition of "effort" focuses on collecting participants for each phase. They propose the use of error-sensitive indicators to identify when a phase has reached capacity and no more participants need to be recruited for that phase. This notion of phase capacity could extend to reaching a stable estimate of a participant's survey-answering, and no more questions need to be asked.

2.2.2. Adaptive Treatment Strategies

In the field of medical statistics, adaptive treatment strategies (also called dynamic treatment regimes) continually adjust treatments, according to decision rules, depending

on an individual's responses to previous treatments as well as characteristics of the patient (Collins et al. 2007). This technique contrasts with the research standard of randomized controlled trials but more closely matches real-world practice of medical intervention (since, when a treatment fails for a particular patient, that patient is reassigned to a new treatment, based on how they reacted). Adaptive treatment strategies are targeted for an *individual*, rather than basing future treatment decisions on outcomes of previous patients.

The design of the sequential multiple assignment randomized (SMAR) trial (Murphy 2005) chooses a decision to make at each point according to what action will maximize the expected treatment outcome, given past information that has occurred. SMAR trials randomize individuals to different treatments at each decision time point.

Adaptive treatment strategies have been applied to treat depression, with the STAR*D (sequenced treatment alternatives to relieve depression) treatment (Rush et al. 2004), in which patients who did not respond to less-intensive therapies were randomly assigned to more intensive treatments at higher levels; to treat schizophrenia, with the CATIE (clinical antipsychotic trials of intervention effectiveness) design (Stroup et al. 2003), a three-phase study where patients were randomly assigned to new treatments at successive phases if they did not respond to earlier treatments; to treat advanced prostate cancer (Wang et al. 2012) by randomizing nonfavorably-responding patients to untried chemotherapy treatments at eight-week intervals, up to four times; and many other medical settings (e.g., smoking cessation (Collins et al. 2005), pediatric generalized anxiety disorders (Almirall et al. 2012), and mood disorders (Lavori et al. 2000; Kilbourne et al. 2014)).

2.2.3. Adaptive Testing

For tests that measure ability or aptitude, adaptive testing selects test questions based on the respondent's answers to previous questions. The goal is to measure the examinee's achievement accurately, without making the examinee answer too many questions. Adaptive tests have been shown to be as reliable and valid as conventional tests (with static question orders), while reducing test length up to 50% (Weiss 1982). Unlike classical test theory, which assumes all questions equally indicate an assessment outcome, item response theory (IRT) (Lord 1980) considers *individual* test questions through an item response function, the probability of a correct answer by an individual at a particular skill level θ. The item response function has three parameters: the pseudo-chance score level (how easy it is to guess the correct answer), item difficulty (how hard it is to answer the question), and discriminating power (how much the skill level influences question response). According to Weiss (1982), an IRT-based adaptive testing framework has these three components: (1) a way to choose the first item to ask, (2) a way to score items and choose the next item to ask during test administration, and (3) a way to choose to end the test, based on an individual's performance.

Weiss and Kingsbury (1984) introduce adaptive mastery testing to assess a student's achievement level $\hat{\theta}$, specifically how the estimated achievement level compares to a "mastery level," θ_m. At each time point, a question is selected which gives the maximum information at the student's current estimated mastery level and asked. As the student answers questions, the estimate $\hat{\theta}$ is updated, along with a confidence interval. Once the confidence interval for $\hat{\theta}$ no longer includes θ_m, the test is finished and the student's mastery level is assigned as sufficient or not (depending if θ_m lies above or below the

confidence interval for $\hat{\theta}$). Kamakura and Balasubramanian (1989) take a similar approach to adaptive question selection for personality trait assessment, but they terminate the process once a minimum standard error is achieved or a minimum number of questions have been asked. These two examples assume dichotomous items (i.e., responses were coded as "correct"/"incorrect" in achievement testing (Weiss and Kingsbury 1984) and "yes"/"no" to responses for personality scoring (Kamakura and Balasubramanian 1989)). Singh et al. (1990) extend the maximum-information question selection criterion to handle graded responses (e.g., Likert scales) and apply this adaptive method to a questionnaire to assess consumer discontent; they achieve comparable estimates as the full-item model at a 33% reduction of items.

More recently, IRT-based adaptive testing has been used for diagnoses of mental health disorders through patient questionnaires, for measuring both unidimensional (e.g., (Gardner et al. 2004; Fliege et al. 2005)) and multidimensional (e.g., (Gibbons et al. 2008 2012; Achtyes et al. 2015)) constructs (see overview in (Gibbons et al. 2016)). For example, Gibbons et al. (2008) use a bifactor IRT model to assess patients' severity of mood and anxiety disorders. Their computerized adaptive testing method yields similar scores as the full-scale version, at a 95% reduction of items. Montgomery and Cutler (2013) have also used IRT-based adaptive testing, but for public opinion surveys. In an empirical study using adaptive testing to measure respondents' political knowledge, the authors found that the adaptive testing approach could produce more accurate measurements than traditional test administration, at a 40% reduction in questionnaire length.

2.2.4. Test-Time Feature Selection

In the case where survey collection is targeted toward the goal of providing the user with a personalized prediction (e.g., for energy consumption), at test time, the goal is to make a prediction on a new example. Making a prediction on a test instance requires gathering values for features (i.e., predictor variables), which can be costly, especially if it requires cooperation from users who might stop before completion. In this case, strategically ordering questions asked (based on previous answers) can get the most useful information first, while providing predictions on partial information. This way, people receive meaningful predictions without spending much time or effort answering questions.

The test-time feature ordering problem resembles active learning, which assumes *labels* (i.e., response variables) are expensive. Active learning algorithms strategically select which unlabeled points to query to maximize the model's performance (using both labeled and unlabeled data) while minimizing the cost of data collection (Cohn et al. 1996). Test-time feature ordering has a similar goal of making accurate predictions while keeping data collection costs low. However, rather than choosing an *example to be labeled*, test-time feature ordering chooses a *single feature to be entered* (by asking the user a question).

He et al. (2012) consider the setting of test-time feature selection, where all features are available for training, and at test time they want an instance-specific subset of features for prediction, trading off feature cost with prediction accuracy. They formulate dynamic feature selection as a Markov decision process (MDP). The policy selects a feature to add; the reward function reflects the classifier margin with the next feature, penalized by the cost of including that feature. However, this method does not make sequential predictions,

and instead only chooses whether to keep getting features or to stop and make a final prediction. Karayev et al. (2012) also take an MDP approach to classify images with a framework they call "timely object recognition," which sequentially runs detectors and classifiers on subsets of the image and uses previous results to inform the next action to take, while providing the best object recognition in the available runtime.

Another related area is personalization, where a system suggests items from user preferences. However, the cold-start problem makes it hard to give recommendations at first, when the system knows nothing about the user, including which questions to ask. Sun et al. (2013) present a multiple-question decision tree for recommendation, where each node asks users for opinions on several movies, rather than just one. This model lets users sooner provide information about movies they have seen, but it is designed to minimize the number of questions and not the amount of effort required to answer questions of varying difficulty.

Most work in test-time feature ordering does not consider the situation of providing predictions with partial information *as* questions are answered, nor does it address the issue of giving measures of prediction quality to users.

3. Case Studies

In this section, we consider two case studies that illustrate how dynamically ordering questions in a survey can be beneficial to survey respondents and data collectors. The first uses the Residential Energy Consumption Survey (RECS) (U.S. Energy Information Administration 2009) to predict energy costs for prospective tenants. We use prediction quality to guide question ordering, trading off how much a new feature is expected to increase the certainty of the next prediction against the effort the respondent must exert to answer a question. We find that respondents can get high-quality predictions at only 21% of the cost of asking the full set of questions. The second case study uses the Survey of Income and Program Participation (SIPP) (U.S. Bureau of the Census 2014b) to gather information about households, mostly related to their use of social programs such as Medicare and Supplemental Security Income. We use the conditional entropy of a respondent's answers to guide question ordering, trading off how much new information a question is expected to give against the likelihood that that question will be answered (measured as the item nonresponse rate). We also simulate survey breakoff and see how a dynamically ordered form can reduce or delay breakoff. We find that the dynamically ordered form can recover enough relevant information to impute unanswered questions, at up to a 57% cost reduction compared to the standard survey question order. Furthermore, even when breakoff rates are fixed between the dynamically ordered form and the standard fixed order form, the dynamic form acquires relevant information to achieve lower bias and variance for survey estimates than the fixed order form.

3.1. Providing Personalized Energy Estimates with the Residential Energy Consumption Survey

In this subsection, we illustrate the concept of dynamic question ordering for prediction with an application of predicting energy consumption for a prospective tenant in a potential home.

Selecting homes with energy-efficient infrastructure is important for renters, because infrastructure influences energy consumption far more than in-home behavior (Dietz et al. 2009). The importance of energy estimates for apartment hunters is twofold. First, since renters often cannot make infrastructure upgrades for efficiency in a property they do not own, they need to know upfront the expected costs of living in a rental unit. Second, 30% of the U.S. population rent, and renters move on average every two years (U.S. Bureau of the Census 2013). Therefore, renters can potentially choose improved infrastructure more frequently than homeowners can make costly upgrades.

Personalized energy estimates can guide prospective tenants toward energy-efficient homes, but this information is not readily available. Utility estimates are not typically offered to house-hunters, and existing technologies like carbon calculators require users to answer (prohibitively) many questions that may require considerable research to answer. For the task of providing personalized utility estimates to prospective tenants, we present a cost-based model for feature selection at training time, where all features are available and costs assigned to each feature reflect the difficulty of acquisition. At test time, we have immediate access to some features but others are difficult to acquire (costly). In this limited-information setting, we strategically order questions we ask each user, tailored to previous information provided, to give the most certain predictions while minimizing the cost to users. During the critical first ten questions that our approach selects, prediction accuracy improves equally to fixed-order approaches, but prediction certainty is higher. We considered this application as one short example for a framework of prediction-guided feature selection in (Early et al. 2016). Here, we include details on the statistical models and methods used to make predictions with partial information and choose a question ordering.

3.1.1. Introduction

Since energy consumption depends on home infrastructure (e.g., square footage) and occupant behavior (e.g., preferred temperature), we can learn the relationship between these features and energy consumption through established data sets, like the Residential Energy Consumption Survey (RECS) (U.S. Energy Information Administration 2009). Some information can be extracted automatically from online rental advertisements, while other information must be provided by prospective tenants at various costs (for example, the question of how many windows a home has requires more effort to answer than how many people will live there). To develop a predictive model of energy consumption at training time, we begin with the extractable (i.e., "free") features in a regression model to predict energy usage and use forward selection to add a subset of the costly features.

After learning this predictive model on the training data set, the main problem lies in how to make a prediction on a new test instance. Initially, only a subset of the features (the free features) in the model is available, and asking users for each unknown value incurs a cost, depending on how hard it is to provide that feature. Our dynamic question-ordering algorithm (DQO) chooses the best question to ask next by considering which feature, if its value were known, would most reduce uncertainty, measured by the width of the prediction interval, with a penalty term on that feature's cost.

3.1.2. Method

Training Time: Cost-Aware Feature Selection A greedy approximation to feature selection, forward selection starts with an empty feature set and, at each iteration, adds the feature that minimizes error (Harrell 2001; Tropp 2004). For this analysis, we started with the free features (rather than no features, as in classic forward selection) and minimized leave-one-out cross-validation error with linear regression to add higher-cost features.

Test Time: Cost-Effective Dynamic Question Ordering After learning regression models for energy usage on the selected features from our training set, we want to make a prediction on a new test point, where initially only some features of the model are available. Our approach considers a trajectory of prediction intervals as a user provides information. A prediction interval consists of a lower and upper bound such that the true value lies in this interval with at least some probability (Weisberg 2014). Prediction interval width corresponds to prediction uncertainty: a wider interval means less confidence. We select as the optimal next question the one whose inclusion most reduces the expected prediction interval width; that is, it most reduces the expected uncertainty of the next prediction.

In this problem, there are features that are unknown (not yet supplied by the user). We use k nearest neighbors (kNN) (Cover and Hart 1967) to supply values for unanswered features in vector $x \in \mathbb{R}^d$. For each unknown feature f, we find the k data points in the training set $X \in \mathbb{R}^{n \times d}$ (n samples, each d-dimensional) that are closest to x, along the dimensions $\mathcal{K}$ that are currently known. Then we estimate x_f as z_f, the mean or mode, as appropriate, of feature f in the k nearest neighbors (depending whether the feature is continuous or discrete).

Because these z values estimate unknown features of x, we use the measurement error model (MEM) (Fuller 2009) to capture error associated with estimated features. Unlike traditional regression models, MEMs do not assume we observe each component x_f exactly. There is an error δ_f associated with the estimation:

$$z_f = x_f + \delta_f, \quad \text{where } \mathbb{E}[\delta_f | x_f] = 0.$$

Prediction $\hat{y}$ still depends on the *true, unobserved* value x:

$$\hat{y} = \hat{\beta}^T \bar{x} = \hat{\beta}(\bar{z} - \bar{\delta}),$$

where $\hat{\beta} \in \mathbb{R}^{d+1}$ is the parameter vector learned on the training set X (recall all feature values are known at training time). The notation $\bar{x}, \bar{z}, \bar{\delta}$ means vectors x, z have a 1 appended to them and δ a 0 to account for the constant term in the regression. Let $\bar{X}$ extend this notion to the training matrix: $\bar{X} = [1^n X]$.

We can calculate a $100(1 - \alpha)\%$ prediction interval for a new point z as

$$\hat{y} \pm t_{n-d-1;\alpha/2} \sqrt{\hat{\sigma}^2 \left(1 + \bar{z}^T (\bar{X}^T \bar{X})^{-1} \bar{z} + \bar{\delta}^T (\bar{X}^T \bar{X})^{-1} \bar{\delta}\right)}, \tag{1}$$

where the $\bar{\delta}^T (\bar{X}^T \bar{X})^{-1} \bar{\delta}$ term accounts for error from estimated features and $t_{n-d-1;\alpha/2}$ is the value at which a Student's t distribution with $n - d - 1$ degrees of freedom has cumulative distribution function value $\alpha/2$. We can estimate δ from training data by

calculating the error of predicting each feature with kNN, from the other features. We also estimate $\hat{\sigma}^2$, the regression variance, from training data.

Then, we cycle through each candidate feature f and compute the expected prediction interval width $\mathbb{E}[w(f)]$ for asking that feature next, over each value r that feature f might take on from its range of potential values R:

$$\mathbb{E}[w(f)] = 2 \cdot t_{n-d-1;\alpha/2} \sum_{r \in R} p(z_f = r) \times \sqrt{\hat{\sigma}^2 \left(1 + \bar{z}_{f:=r}^T (\bar{X}^T \bar{X})^{-1} \bar{z}_{f:=r} + \bar{\delta}_{f:=0}^T (\bar{X}^T \bar{X})^{-1} \bar{\delta}_{f:=0}\right)}, \tag{2}$$

where $p(z_f = r)$, the probability that the fth feature's value is r, is calculated empirically from the training set, and the notation $\bar{u}_{f:=q}$ means the f-th component of u is replaced with the value q.

Including the feature that attains the narrowest expected prediction interval width $\mathbb{E}[w(f)]$ will reduce the uncertainty of our prediction more than any other feature. This approach allows incorporation of feature cost into the question selection, by weighting the expected prediction interval width against the cost of acquiring the feature:

$$f^\star = \arg\min_f \left(\mathbb{E}[w(f)] + \lambda c_f\right), \tag{3}$$

where c_f is the cost of feature f and $\lambda \in \mathbb{R}$ trades off feature cost with reduced uncertainty. A high-cost feature might not be chosen, if another feature can provide enough improvement at lower cost. We ask for this information, update our vector of known data with the response (and estimate the unknown features again, now including the new feature in the set for kNN prediction), and repeat the process until all feature values are filled in (or the user stops answering).

More generally, this algorithm can be seen as a framework that makes predictions on partial information and selects which feature to query next by (1) estimating values for unknown features (here with kNN) and (2) asking for the feature that will most reduce the expected uncertainty of the next prediction (here measured by prediction interval width). With this approach, we strategically order questions, tailored to previous information, to give accurate predictions while minimizing the user burden of answering many or difficult questions that will not provide a substantial reduction in prediction uncertainty.

3.1.3. Data

The Residential Energy Consumption Survey (RECS) contains information about home infrastructure, occupants, and energy consumption. We can use this data set to learn relationships between household features and energy consumption to predict energy usage for prospective tenants. The most recently released RECS was a nationally representative sample of 12,083 homes across the U.S. (U.S. Energy Information Administration 2009). For each household, RECS records fuel consumption by fuel type (e.g., electricity, natural gas) and around 500 features of the home (e.g., age of refrigerator, number of occupants).

Defining Feature Costs In our problem setting, features have different costs of being obtained, and we want to build models and make predictions that leverage features cost-effectively. Some information is easily found in the rental listing (e.g., number of bedrooms) and is therefore "free," while other information requires asking users. For example, the number of windows does not appear in listings and would require a prospective tenant to visit each site. Consequently, this question has high cost. Other useful features relate to occupant behavior (e.g., preferred temperature). These questions likely remain constant for each user across homes and therefore require asking only once and are cheaper. We categorize feature costs as "extractable/free" (can be automatically extracted from rental listings), "low" (occupant-related; require asking only once), and "high" (unit-related; must be answered once for each apartment and may require a site visit). Table 1 lists the information used for extractable features and how often it appears in Rent Jungle, a company that scrapes rental listings from the Internet. These "free" features appear in the majority of listings on Rent Jungle. In our experiments, we assume all homes have a rental advertisement with all of this extractable information. If a home does not have an advertisement (or its advertisement is missing information), the missing information can be included in the list of features to acquire from the user, with no change to our method.

3.1.4. Experimental Validation

We validated our test-time feature ordering approach on the RECS data set for predicting household electricity and natural gas consumption. We restricted our analysis to homes in the same climate zone as our planned deployment location in Pittsburgh, a subset of 2,470 households in climate zone 2. We used 90% of these homes for training and the remaining 10% for testing. The training set was further subdivided into feature selection and cross-validation subsets. We trained and tested separate models for predicting electricity consumption (on all homes in our climate zone) and for predicting natural gas consumption (on the 75% of homes that use natural gas) with forward feature selection (Harrell 2001; Tropp 2004). Due to the similarity of the results from electricity prediction and natural gas prediction, we show here results from only electricity prediction.

Test Time: Cost-Effective Dynamic Question Ordering After learning regression models for electricity and natural gas prediction on the training set, we want to make predictions of energy usage for a new test point. We use our dynamic question-ordering framework (DQO) to make sequential predictions with partial, evolving information. To apply DQO,

Table 1. The features we define as extractable (i.e., "free") appear in most of the listings on Rent Jungle. Geographic features associated with the city, zip code, or state include climate zone and whether the area is urban or rural, among others.

Feature	Presence in Rent Jungle database (%)
Number of bedrooms	85
Number of full bathrooms	57
Studio apartment	85
City or zip code	99
State	100

we first used the training set to choose parameters for imputing unknown features (k, for k-nearest neighbors) and to estimate the measurement error δ. Then, we simulated question-asking on RECS to evaluate the performance of DQO for test-time feature ordering.

Parameter Selection and Estimation We used the training set to choose $k = 100$ for imputing the values of features that have not yet been asked, based on the prediction performance of kNN for the higher-cost features. Then, we estimated the measurement error δ_f for each feature as the error from kNN on the training set.

Simulating the Question-Asking and -Answering Process With RECS We simulated the process of asking and answering questions on the testing subset of RECS by hiding the values for features that were not yet known. After we used DQO to choose a feature to acquire, we "asked" this question and unveiled its value once it was "answered."

Evaluating DQO Performance with Prediction Certainty, Error, and Cost We evaluated the performance of our cost-effective *DQO* algorithm in making sequential predictions with partial, evolving information on a held-out test set. For comparison, we implemented several baselines. The *Random* algorithm chooses a random question ordering for each sample, the *Fixed Decreasing* algorithm asks questions in decreasing order of feature measurement error δ_f (identical ordering for all samples), and the *Fixed Selection* algorithm asks questions in the order of forward selection in the training phase (also identical for all samples). Finally, the *Oracle* chooses the next best feature according to the minimum true prediction interval width (calculated on the test sample using true feature values, rather than the *expected* width as in Equation 3). We tested two versions of DQO and oracle: ordering additional features *without cost* and *with cost* (implemented as $\lambda = 0$ and $\lambda > 0$). The two *Fixed* algorithms illustrate the performance of a fixed order of questions, applied to all respondents. The *DQO* algorithms show the benefits of adapting question order to the individual respondent (compared to the baseline *Random* order, where each respondent gets a different ordering, but not based on responses or prediction quality). Finally, the *Oracle* algorithms demonstrate the optimal performance of a personalized question order, since they rely on knowing the true feature values, even before they are provided by the respondent.

We calculated several metrics related to the trajectory of prediction performance and cost, for orderings given by the algorithms: *DQO* and *Oracle with* and *without cost, Random, Fixed Decreasing*, and *Fixed Selection*. We summarized prediction performance with the width of the current prediction interval (prediction *certainty*) and the absolute value of the difference between the current prediction and the truth (prediction *error*). We also measured the cumulative cost of all features asked at each step (prediction *cost*). Table 2 summarizes the metric trajectories as areas under the curve—smaller values are better because they mean the algorithm spent less time in high uncertainty, error, and cost.

Certainty Metrics For certainty, we calculated widths of 90% prediction intervals as features were answered. Since narrower prediction interval widths correspond to more certain predictions, we expect DQO interval widths to be less than those of the baselines, particularly in the early stages. Figure 1a plots the *actual* prediction interval widths as questions are asked (calculated with Equation 1, using the true known feature values and imputed values for unknown features), averaged across the test data set, for the question

Table 2. Areas under the curve for the certainty, error, and cost metrics from various methods, for electricity prediction: smaller values mean the algorithm spent less time in high uncertainty, error, and cost.

Method	*Interval width*	$\lvert y - \hat{y} \rvert$	*Cost*
DQO without cost	42.43	12.06	1212.91
DQO with cost	42.44	12.53	1120.50
Random	42.53	13.18	1213.30
Fixed decreasing	42.46	11.85	1233.50
Fixed selection	42.47	11.45	1190.50
Oracle without cost	42.35	14.62	1222.91
Oracle with cost	42.39	13.63	1120.50

sets from the seven orderings. The DQO sets result in the narrowest (or near-narrowest) prediction intervals (i.e., most certain predictions), compared to the baselines, with improvements most notable in the first ten questions answered – the situation when users do not answer all the questions.

Error Metrics For error, we calculated the absolute value of differences between the midpoint of the 90% prediction interval and truth as questions are answered, plotted in Figure 1b. Because this metric compares a point prediction to the true value, error is still incurred for prediction intervals that include the true value (as 90% of them will, by construction), when the true value is not the exact midpoint of the range. For all orderings, predictions approach the true value as questions are answered. Once about ten questions have been asked, *DQO with cost* reaches similar performance as *DQO without cost* and the fixed order baselines (*Fixed decreasing* and *Fixed selection*).

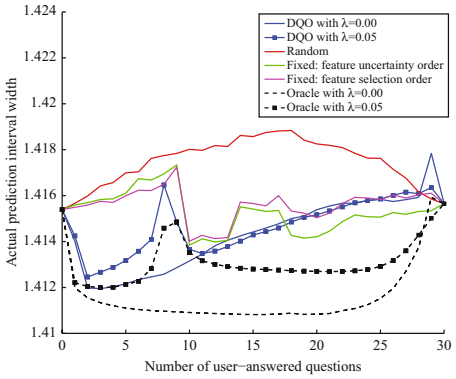
Cost Metrics Progressive total feature costs as features are asked and their true values are used in the models are plotted in Figure 1c. Cumulative feature costs are lower for orderings that penalize feature cost (*DQO, Oracle with cost*), with cost decreasing as the penalty on cost λ increases. The other orderings have similar cost trajectories to each other.

Overall, these metrics show that our test-time DQO approach quickly achieves accurate, confident predictions. By asking around ten questions, DQO (with and without cost) reaches similar accuracy as the fixed order baselines, but the sequential predictions by the fixed orderings are less confident than DQO until about 20 questions have been asked.

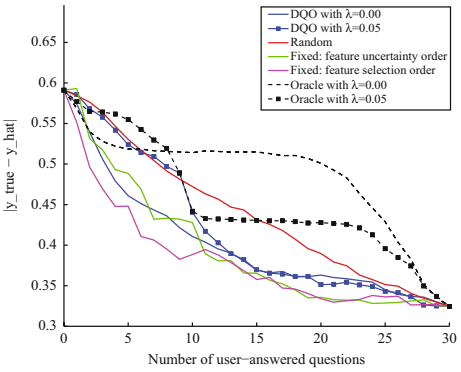
Figure 1d shows how frequently the oracle asked each feature in each position across test instances. Most features are chosen fairly uniformly at each position in the question ordering. This indicates that there is no single best orderto ask questions across all households, which is why the dynamic question-ordering process is so valuable.

3.1.5. Limitations

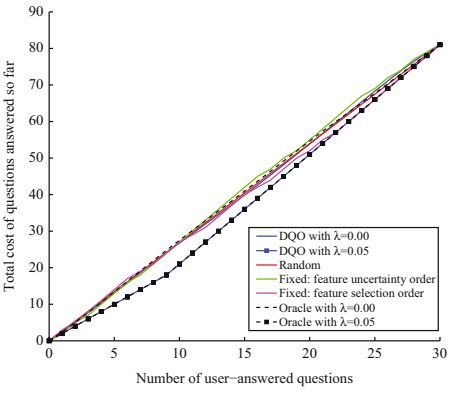
Currently, our DQO algorithm assumes that (1) users are able to answer the next question we ask and (2) their answers are accurate. However, situations could arise where these assumptions do not hold. For example, in the energy prediction task, a prospective tenant may be interested in getting personalized energy estimates for a home before visiting – they could still answer occupant-related features. DQO can be easily extended to this case



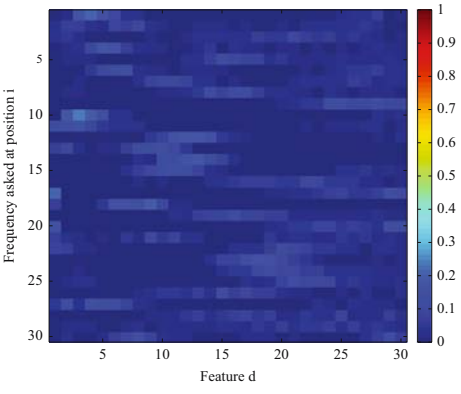
(a) Prediction interval widths as questions are asked: DQO results in more certain predictions (i.e., lower prediction interval widths) than the baseline orderings.



(b) Mean absolute error as questions are asked: DQO results in similarly-correct predictions as baselines.



(c) Total feature costs as questions are asked: the tradeo parameter λ influences when expensive features are included.



(d) The oracle chose to add features fairly uniformly across test samples, shown here as the low frequency each feature was asked in each position question orders.

Fig. 1. Results from dynamically ordering questions for test-time electricity prediction.

by offering users a "don't know" option for answering questions and removing unknown features from consideration in later iterations. Breaking the second assumption, that user answers are accurate, would allow people to give estimates for features (e.g., refrigerator size by looking at pictures in the rental listing). Incorporating this element into DQO would require a way to estimate error associated with user-provided feature estimates.

Furthermore, we have not yet tested the DQO question-asking and prediction-providing process with human users. We hypothesize that giving users estimates from partial information will motivate them to continue answering questions to receive more accurate personalized predictions. On the other hand, once the prediction interval width is small enough or stable enough, users may no longer see the value in continuing to answer questions and will stop.

3.1.6. Conclusion

Providing personalized energy estimates to prospective tenants with limited, costly information is a challenge. Our solution uses an established data set to build cost-effective

predictive models and, at test time, dynamically orders questions for each user. At test time, when we want to make a personalized estimate for a new renter-home pair, we present a cost-effective way to choose questions to ask a user about their habits and a rental unit, based on which feature's inclusion would most improve the certainty of our prediction, given the information we already know. Our experiments show that, for predicting energy consumption, we achieve prediction performance that is equally accurate to, but more certain than, two fixed order baselines by asking users only 21% of features (26% of the cost of the full-feature model).

3.2. General Data Collection with the Survey of Income and Program Participation

A prediction-guided approach to dynamic question ordering (Subsection 3.1) is only reasonable when the goal of information gathering is to make a prediction. Now we consider how to dynamically order questions in a way that most characterizes the respondent so that if they do drop out of the survey, imputed values for their unanswered questions will be accurate.

3.2.1. Introduction

The Survey of Income and Program Participation (SIPP), conducted by the U.S. Census Bureau, collects data on income, employment, and social program participation and eligibility from households (U.S. Bureau of the Census 2014b). The SIPP is designed as a longitudinal national panel survey, where each panel is a representative sample of 14,000 to 52,000 households, contacted yearly for three to five consecutive years. Each household interview is conducted in person, via a computer-assisted personal interviewing (CAPI) instrument, and aims to get self-reports from all household members at least 15 years old. In addition to demographic information, interviews ask respondents for their participation in various social programs, financial situation, and employment status, in the previous calendar year. The chief goal of SIPP is to understand household program eligibility and participation and to assess the effectiveness of social programs like Supplemental Security Income, Supplemental Nutrition Assistance Program, Temporary Assistance for Needy Families, and Medicaid.

Since responses to survey questions tend to be related, we can use information-theoretic concepts like entropy to measure the information content of a new question, taking into account what we already know about the respondent from their previously provided responses.

3.2.2. Method

To measure the amount of information expected to be contained in answers to potential questions that might be asked next, we use *conditional entropy* as the measure of utility (Cover and Thomas 2012). Entropy measures the information in a random variable X. If this random variable can take on one of D possible values, its entropy is defined as

$$H(X) = -\sum_{x \in X} p(x) \log p(x). \tag{4}$$

Conditional entropy measures the information in a random variable Y, when the value of another variable X is already known:

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x). \tag{5}$$

When iteratively selecting which question to ask next, we want to maximize conditional entropy while keeping costs low:

$$q^{\star} = \arg\min_{q} \left(-H(q|X) + \lambda c_q \right), \tag{6}$$

where q represents a question, X is the set of questions that have already been answered, c_q is the cost of question q, and λ is a tradeoff parameter that controls how heavily cost is weighted in this question selection rule.

3.2.3. Data

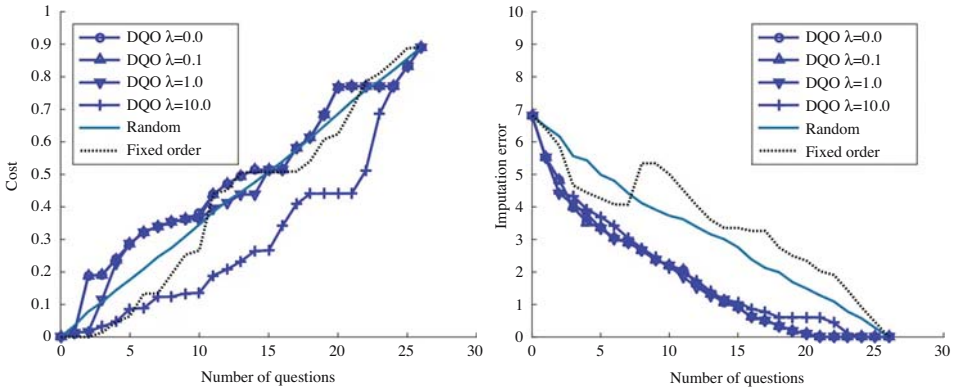
In our experiments, we use the SIPP Synthetic Beta (SSB), a synthetic data set created by first multiply-imputing missing values in the SIPP Gold Standard File and then multiply-imputing replacement values for actual responses to preserve privacy (Benedetto et al. 2013). The SSB combines SIPP data with administrative records from sources like the Internal Revenue Service. These data from administrative records are not asked of a respondent and can be considered initializing information when selecting questions to ask.

Defining Question Costs In this case study, we use item nonresponse rates as a proxy for question burden. We calculate item nonresponse rates as the fraction of respondents for whom a value to a particular survey item had to be imputed, using publicly-available SIPP data (http://thedataweb.rm.census.gov/ftp/sipp_ftp.html). We consider data from administrative records in the SSB to be "free," since no household had to answer questions to get those values.

3.2.4. Experimental Validation

In the experiments in this subsection, we use conditional entropy to measure how much information a new question will give, in light of what is already known about the respondent. We order questions and calculate performance metrics for the DQO-ordered questions with a variety of cost penalties λ, a random order selected for each respondent, and a fixed order baseline according to the order in which these questions are asked in the SIPP interview. At each point in the question-asking process, we calculated the cost and imputation error. Cost is defined as the sum of the costs (nonresponse rates) for all questions asked thus far. Imputation error is defined as the number of incorrect imputed values for yet-unanswered questions. We impute values for unanswered questions by using k-nearest neighbors to find the points in the training data set nearest to the current respondent and predicting unknown values as the mode of those values among the nearest neighbors. This is not the method used for imputation in the SIPP (U.S. Bureau of the Census 2014b).

Figure 2 shows the costs (Figure 2a) and imputation errors (Figure 2b) as questions are asked in SIPP for the DQO-ordered sets, a random question ordering, and the fixed order baseline.



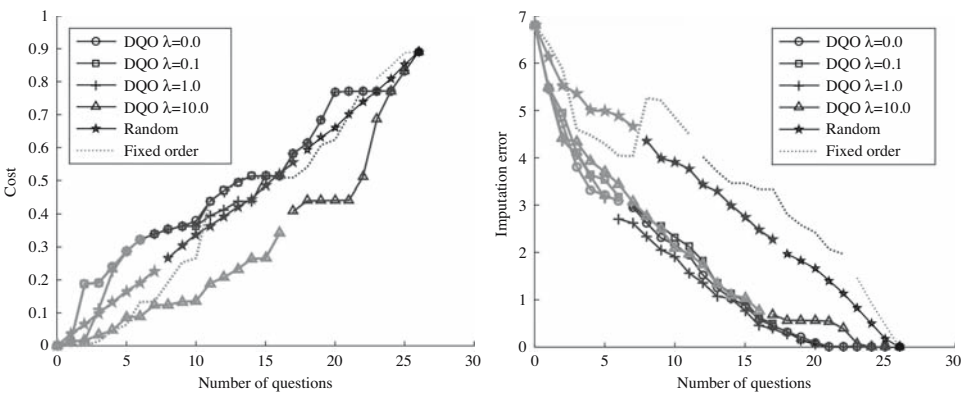
(a) As the cost penalty λ increases, the cost savings from DQO increase, compared to the fixed order baseline.

(b) DQO has lower imputation error than the fixed order baseline.

Fig. 2. Charts showing impact on cost and imputation error (y axis) of λ and number of features, for DQO (solid lines) and the fixed order baseline (dashed lines).

Simulating Breakoff in Survey Collection Next, we simulated survey breakoff at the presentation of each question by randomly choosing if a respondent would "drop out" of the survey, with probability according to the cost of the current question. If the person did drop out, their responses to that question and all remaining questions would be left blank. In this simulation, we assume that item nonresponse comes only from these breakoffs (i.e., there are no "skipped" questions).

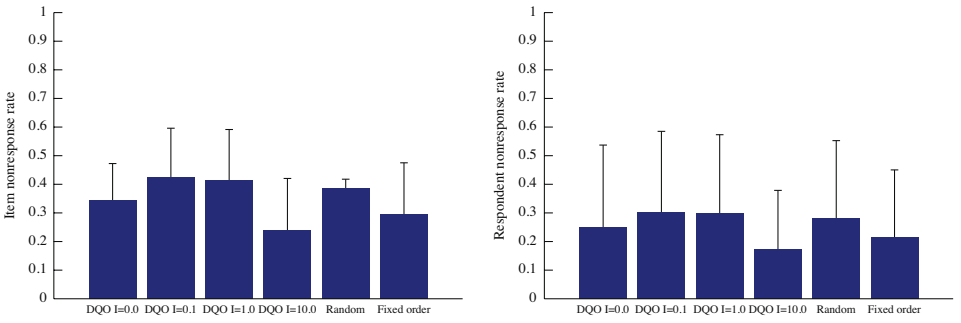
Figure 3 plots the cost (Figure 3a) and imputation error (Figure 3b) trajectories for DQO with several cost penalties λ, the random ordering for each respondent, and the fixed order baseline. The colors in the plot represent when at least 75% (light gray), 50–75% (gray), and fewer than 50% (black) of respondents answered that number of questions. DQO consistently attains lower imputation error than the fixed order baseline and the random



(a) As the cost penalty λ increases, the cost savings from DQO increase, compared to the fixed order baseline.

(b) DQO has lower imputation error than the fixed order baseline.

Fig. 3. Charts showing impact on cost and imputation error (y axis) of λ and number of features, for DQO (solid lines) and the fixed order baseline (dashed line). The color of each segment reflects how many respondents reached that far in the survey: at least 75% answered questions in light gray, 50–75% answered questions in gray, and fewer than 50% answered questions in black.



(a) The cost penalty λ affects the item nonresponse rates in DQO.

(b) The cost penalty λ also affects the respondent nonresponse rates in DQO.

Fig. 4. Plots showing nonresponse rates for each question (fraction of respondents who did not answer a question, due to dropout) and each respondent (fraction of questions each respondent did not answer when they broke off): lower values are better.

ordering. The imputation error is similar for all DQO orderings, regardless of cost penalty. The cost penalty does affect the cost trajectory, with higher cost penalties resulting in lower cost trajectories.

Figure 4 illustrates the nonresponse rates for the different methods: Figure 4a plots the average nonresponse rate for each question (fraction of respondents who answered each question before dropping out), and Figure 4b plots the average nonresponse rate for each respondent (fraction of questions each respondent answered before dropping out). For both types of nonresponse, DQO performs similarly to the fixed order baseline when $\lambda = 0$, slightly worse when $\lambda = 0.1,1.0$, and better when $\lambda = 10$. Figure 5 shows individual item

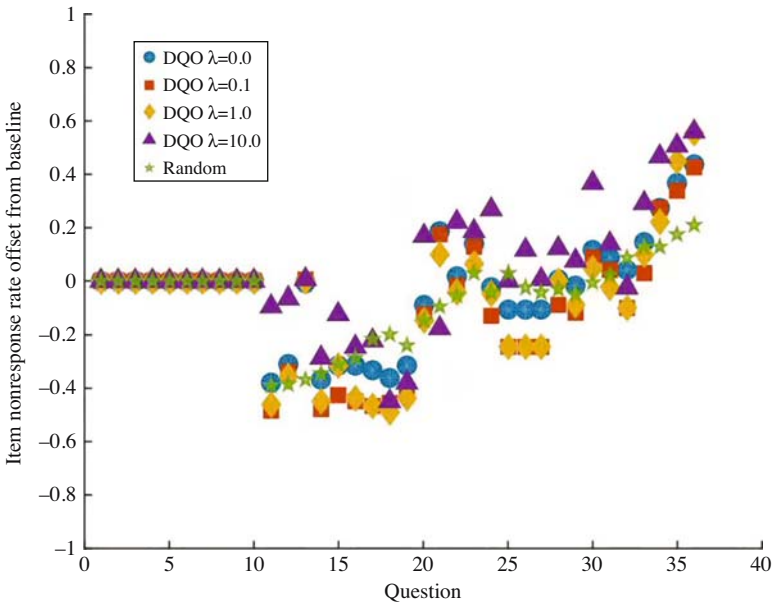


Fig. 5. The offset of item nonresponse rates from the baseline's: positive values mean a method has lower item nonresponse rate than the fixed order baseline. The first ten items come from administrative records and are therefore "free" (not asked during the interview).

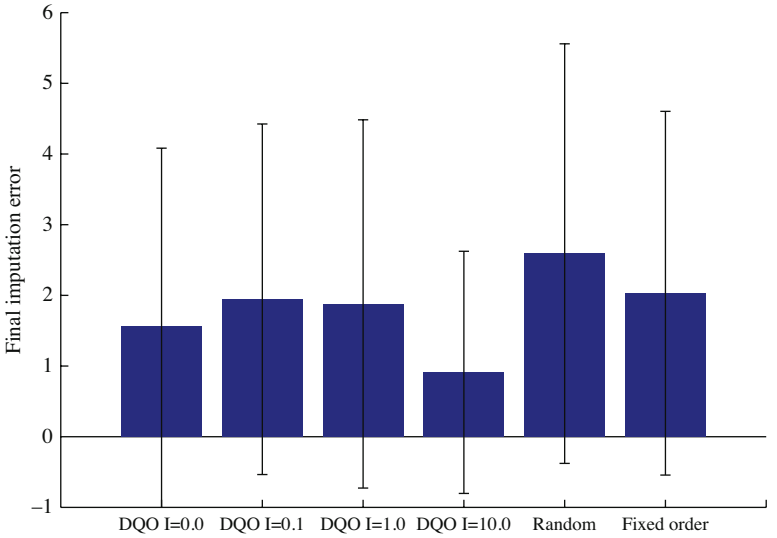


Fig. 6. Final imputation error (number of incorrectly-imputed values once the respondent finishes (including by dropping out) the survey: lower values are better.

nonresponse rates, compared to the fixed order baseline. This plot illustrates how DQO can achieve lower nonresponse (i.e., higher values in the plot of "baseline item nonresponse rate" – "DQO item nonresponse rate") for later items in the survey, since those items might be asked earlier and therefore are less likely to suffer from dropout.

Figure 6 plots the final imputation error for each method. Final imputation error is the total number of incorrectly-imputed values once a respondent stops providing answers. For this metric, all variants of DQO outperform the fixed order baseline. Even though some of the DQO orderings have higher nonresponse rates than the fixed order baseline (Figure 4), DQO tends to choose the most informative questions near the beginning so imputation quality does not suffer even when a respondent drops out before completing the survey.

Bias and Variance of Survey Estimates To compare performance in terms of the quality of final survey estimates (e.g., population means), we set the DQO orderings and the fixed orderings to have 16% breakoff rates (the average breakoff rate for web surveys (Manfreda and Vehovar 2002)). We ran multiple imputation to get multiple sets of imputations for the values missing due to dropout (Rubin 2004). Then we calculated the bias and variance of the resulting survey estimates. Table 3 shows that the estimates from DQO have bias and variance lower than or similar to those from the standard SIPP order.

3.2.5. Limitations

The SIPP Synthetic Beta (SSB) data set has several downsides. First, household structure is not preserved – each respondent is treated as a separate entity. Though there are spouse links for female-male married couples, this information is insufficient to recover a household structure which may include same-sex married couples, unmarried partners, children, other relatives, and other household members. Second, because the SSB is

Table 3. Bias and variance of final survey estimates for SIPP order and DQO orders. Bias was calculated as the (absolute value of the) difference between the calculated survey estimates (which include imputed values) and the true values (from the completedata). Variance was calculated as the variance of survey estimates, across the multiply-imputed data sets. For both metrics, means and standard deviations are given across the survey variables.

Method	Bias	Variance
SIPP order	0.063 ± 0.194	0.013 ± 0.056
DQO, $\lambda = 0$	0.034 ± 0.111	0.005 ± 0.023
DQO, $\lambda = 0.1$	0.047 ± 0.103	0.007 ± 0.031
DQO, $\lambda = 1.0$	0.051 ± 0.146	0.012 ± 0.054
DQO, $\lambda = 10$	0.024 ± 0.063	0.001 ± 0.004

synthetic and fully imputed, it does not have any information on when respondents break off from the survey. However, since we simulated breakoff as proportional to question cost (measured as item nonresponse rate), we expect our breakoff-related results to hold for the case of known breakoff when breakoff probability is the cost metric.

3.2.6. Conclusion

We can capitalize on the relationships among survey responses when dynamically ordering questions to collect the most informative data first, so that if a respondent drops out before finishing the survey, imputation quality will be good. Additionally, when such a method for dynamic question ordering takes breakoff probability into account when ordering questions, breakoffs will be delayed until after the most relevant information is collected. Although this approach means that certain items are more likely to be unanswered (i.e., high-cost items), the dynamic process collects enough low-cost information from the respondent before they break off, so that imputation quality is higher than a fixed order for the questions. Furthermore, even when DQO and standard question orders are fixed to have the same amount of breakoff, DQO collects information such that final survey estimates have bias and variance that are less than or similar to those of estimates from a fixed question order.

4. General Framework for Dynamic Question Ordering in Online Surveys

4.1. Selecting an Optimal Next Question to Ask

To choose an optimal next question to ask in our two case studies, DQO maximized the *utility* of a new question, penalized by the *cost* of that question. For personalized energy estimates using RECS, the measure of utility was prediction certainty, measured by the expected width of the prediction interval, while the measure of cost was the difficulty to the respondent of answering a certain type of question. For survey-taking on SIPP, the measure of utility was conditional entropy, and the cost metric was item nonresponse rates for individual questions.

We can generalize the specifics of these two case studies into a question selection rule that trades off the expected utility of having an answer to a question (or a set of questions)

with the cost of getting the answer:

$$q^{\star} = \arg\min_{q} \left(-\mathbb{E}[U(q)] + \lambda c_q \right), \tag{7}$$

where q is a question or set of questions, $U(q)$ is the utility of q, c_q is the cost of q, and λ is a tradeoff parameter. We want to maximize utility while minimizing costs. Definitions of question utility and cost will vary according tothe application and the purpose of the data collection. This iterative question selection rule allows for both the utility and cost metrics to take into account all information provided up to that point in the survey. For example, if a person already answered the RECS question that they find their home too drafty all the time, asking them next about the insulation of their home would not be very informative (i.e., that question would have low utility), since their home is probably poorly insulated.

4.1.1. Question Utility

Intuitively, question utility captures how "useful" a candidate question is. The exact definition for utility depends on the application (i.e., what do we value as useful?) and the data (i.e., how can we calculate this value?). For example, if the application is a prediction task, we can use the impact a question will have on the prediction quality as a definition for utility. One aspect of prediction quality is *certainty*. The calculation of prediction certainty as a measure of utility will depend on the data and the predictive model being used. If the value to be predicted is continuous, one measure for prediction certainty is the width of the prediction interval, where a wider interval means a less certain prediction. The mathematical definition for the prediction interval width will depend on the predictive model being used. If the value to be predicted is discrete, one measure for prediction certainty is the distance of the sample from the decision boundary. Again, calculation for this measure of certainty (the distance from the decision boundary) will depend on the predictive model.

To be dynamic, these definitions of utility will take into account information already known about the sample. Such information may come from previously answered questions or paradata collected in the survey process. Paradata in particular will be helpful for informing a utility function that captures respondent engagement.

When defining a metric for question utility, it can be challenging to reflect the multiple purposes of a survey in a single utility function. Large-scale government surveys in particular often have multiple federal agencies who care about different questions. Developing a composite utility function that meets the needs of all stakeholders in a multi-purpose survey is important, but also complicated and context-specific, and a topic for continued work.

4.1.2. Question Cost

The cost of a question reflects how "difficult" it is to get an answer for that question. Different applications have different cost measures that are best suited to them. Examples of cost include (1) the amount of resources needed to answer the question (e.g., time, money, battery, effort), (2) the likelihood that the question will not be answered (i.e., item nonresponse rate), (3) the likelihood that the question will cause the respondent to stop the survey (i.e., item breakoff rate), and (4) a combination of multiple types of cost. These

costs can be predefined according to rules (as in the RECS case study) or determined empirically from collected data (as in the SIPP case study). Respondent burden in surveys is multi-faceted, with multiple factors influencing what is ultimately a subjective measure on the respondent (e.g., (Bradburn 1978; Fricker et al. 2014; Yu et al. 2015)). Thus, it is difficult to measure the actual burden a question poses to a respondent. Our case studies in Section 3 used simple proxies for respondent burden. However, this framework is general, and so any more cognitively precise measure can be used, were it available.

When determining question order is deferred to test time, these costs can be *context-dependent*. That is, the current situation of data collection can inform the costs of future questions that might be asked. For example, in a system that makes medical diagnoses and can request tests (e.g., (Ferrucci et al. 2013)), it will be less costly to request an invasive biopsy if a surgery is already scheduled in that area. Context-dependent costs can also be used to address questions that are sensitive to order effects (McFarland 1981). At response time, the cost of a dependent question can be assigned to be infinite when its prerequisite question(s) have not already been asked. Thus, such a question would never minimize the objective in Equation 7 and would never be chosen if the respondent has not already seen questions that need to precede it.

This method can be generalized to order *modules* of related questions, rather than individual questions. Reasons to present questions in modules rather than purely sequentially include (1) presenting related items in a group can reduce the *cognitive burden* required of a respondent to answer the group (e.g., if a set of questions asks the respondent about various aspects of their commute, as the American Community Survey does, it will be easier for the respondent to answer those commute-related questions as a unit rather than scattered throughout the entire questionnaire) (Tourangeau 1984) and (2) imposing a standard order on certain questions that are susceptible to *order effects* (McFarland 1981), in a module, can ensure that all participants understand and answer questions in the same way, even when the order of question modules is determined dynamically.

4.2. *Terminating Data Collection*

A key component of computerized adaptive testing procedures is a criterion for determining when to stop administering test items (Weiss 1982). Typically, the test is terminated once the estimate about the testee is certain enough (e.g., (Weiss and Kingsbury 1984; Kamakura and Balasubramanian 1989)). A similar stopping criterion could be used for prediction-guided DQO (e.g., the RECS case study), but most often surveys aim for complete response. In that case, it is advantageous to ask questions as long as the respondent is willing to answer them. One exception could be for longitudinal surveys, where respondent attrition is partly due to panel fatigue (Laurie and Scott 1999). Terminating a panel survey before completion would reduce respondent burden and could increase willingness to participate in future surveys.

4.3. *Assumptions and Requirements*

Most generally, this approach requires only definitions of question utility and cost and ways to calculate the expected utility of each candidate question given what is known. In

practice, a training set of question responses is needed: measures of utility often depend on statistical properties of samples and estimators, and distributions of question responses need to be known to calculate expected values.

5. Other Potential Survey Applications

In this section, we elaborate on particular problems DQO could solve in other surveys. All surveys in this section are administered in computerized modes in which DQO would be possible. Some surveys, such as the Current Population Survey and the National Crime Victimization Survey, have predictions as goals (identifying employment status or classifying incidents of victimization) and can directly incorporate a prediction-motivated approach to DQO, as in Subsection 3.1. Other surveys, like the American Community Survey (ACS) and National Health Interview Survey, have the broad goal of collecting information on a population. DQO in these surveys would need to focus on maximizing information gain or respondent engagement, calculated from previously provided answers and paradata. National surveys are complicated, due to complex sampling requirements (e.g., oversampling certain populations) and multi-purpose goals (e.g., adding supplemental modules to core surveys).

Typically, statistical agencies already have strategies for handling nonresponse, and the current protocol could influence DQO. For example, the Census Bureau defines several categories of nonresponse in the ACS: households that answer basic demographic and housing questions, but not any detailed person questions, are considered "sufficient partial" responses who are not contacted for follow-up. Households that do not answer the basic questions are "insufficient partial" responses and are contacted for follow-up (U.S. Bureau of the Census 2014a; Clark 2014). Thus, Census might want to apply DQO within sections of the survey: first order questions in the basic subset, to ensure that more people will become at least sufficient partial responses, if not complete. Then, once a household's sufficient partial status is confirmed, the survey can continue with the detailed person questions.

5.1. American Community Survey

The goal of the mandatory American Community Survey (ACS) is to gather complete statistics on the U.S. population, and follow-up with nonrespondents is expensive. Each year 3.54 million households receive mailed surveys to answer between 77 and 347 questions, depending on the number of occupants (U.S. Bureau of the Census 2016). The survey takes, on average, 40 minutes to complete and 54% of homes return theirs (U.S. Bureau of the Census 2014a). The Census Bureau calls nonrespondents for telephone interviews and then samples nonrespondents for home interviews. In 2012 in-person follow-up amounted to 129,000 person-hours per month (Griffin and Nelson 2014). Furthermore, in-person interviews can bias survey results, if nonresponse adjustment weights are not allocated correctly (U.S. Bureau of the Census 2014a). The Census Bureau tested an online survey and found similar data quality for Internet and mail return (Horwitz et al. 2012). Overall response rates were similar, but online surveys had higher item response rates for earlier questions and more blank responses for later questions than paper surveys (Horwitz et al. 2012). Dynamically ordering questions in the online form could

ensure that even if households do not complete the survey, they answer the most informative questions before breaking off.

The online mode for the ACS also collects paradata as respondents complete the survey. These paradata include clicked links (including navigation buttons, responses, help buttons), timestamps, field values, errors, invalid logins, timeouts, logouts (Horwitz et al. 2012). Such paradata could be used to model user engagement, understanding, and willingness to respond, as another component for dynamic question ordering to increase response rate.

5.2. *Current Population Survey*

The Current Population Survey (CPS) is a monthly survey of 60,000 households across the United States, jointly sponsored by the U.S. Census Bureau and Bureau of Labor Statistics (U.S. Bureau of the Census 2006). Selected households are in the survey for four consecutive months, out of the survey for eight months, and then back in the survey for four months. The chief purpose of the CPS is to estimate the U.S. unemployment rate for the past month, and the majority of the official survey is devoted to this task. Respondents answer a battery of questions related to their work status in the past week to determine if they were employed, unemployed, or not in the labor force. There are over 200 questions in the labor force portion of the items. Not all of these questions apply to every household, so the current version of the CPS uses predefined skip patterns to avoid asking irrelevant questions. Augmenting the rule-based skip patterns with dynamic question ordering derived from statistical properties of the respondent could further lower respondent burden.

Various survey sponsors add supplemental question modules to the CPS (e.g., the Tobacco Use Supplement, sponsored by the National Cancer Institute), which may also benefit from dynamic question ordering. The number of supplements is heavily restricted, due to not wanting to overburden respondents with too many questions and detract from the main purpose of the survey – estimating employment rate (U.S. Bureau of the Census 2006). Using dynamic question ordering within or between modules could effectively selectitems to ask of populations of interest, thereby reducing the effective number of questions respondents must answer and increasing the potential for supplemental questions on the CPS.

5.3. *National Health Interview Survey*

The largest U.S. health survey, the National Health Interview Survey (NHIS) is administered in person to about 35,000 households each year (National Center for Health Statistics 2014). The main purpose of the NHIS is to collect health information, at the household, family, and individual levels. Currently, the NHIS uses predefined skip patterns to advance respondents through the survey, but a statistical approach to question ordering could enhance the survey. Like the CPS, the NHIS has supplements to the main survey sponsored by other agencies.

The structure of the NHIS designates one person from a household as the "household respondent" who provides information for all members in the household (even for multi-family households). This type of proxy reporting is more likely to have errors than

self-reports (Sudman et al. 1996), and so a dynamic question-ordering procedure would need to consider the impact of uncertain provided values when choosing which question to ask next.

The NHIS oversamples underrepresented populations, like black, Hispanic, and Asian people, to obtain more precise estimates for these populations (National Center for Health Statistics 2014). With this goal in mind, DQO for the NHIS could consider the accuracy of imputed values for questions that are not yet asked, with thresholds for allowable imputation error. Such thresholds could be population-specific, with minorities' having lower allowable error thresholds, to ensure that more complete data are collected from these populations.

5.4. National Crime Victimization Survey

Every year the U.S. Census Bureau, on behalf of the Bureau of Justice Statistics, administers the National Crime Victimization Survey (NCVS) to 90,000 households (Bureau of Justice Statistics 2014). Occupants of sampled households are interviewed every six months over three years. In interviews, respondents report victimizations in the previous six months.

The NCVS collects detailed information about each incident reported by a respondent to classify incidents into fine-grained categories of crime (e.g., "Robbery – attempted with injury"). The current NCVS design asks a respondent a set of questions regarding each incident they report and uses answers to these questions to classify the crime, rather than directly asking respondents for the crime category (Bureau of Justice Statistics 2014). As such, this survey has a per-individual prediction problem at its core (labeling an incident as a type of crime), like the personalized energy estimate example presented in Subsection 3.1, and could benefit from a similar DQO process, except for classification rather than regression.

DQO could further benefit the NCVS because, especially as households complete the survey multiple times, respondents recognize that reporting an incident results in an extended set of questions. This full questioning takes place for each individual report, including repeat victimizations (e.g., domestic violence). To speed up the interview, participants are likely to underreport victimization, to avoid lengthy subsets of questions for each report. By reducing the number of questions to categorize incidents and using previously provided information to help in question ordering for repeat incidents, DQO could reduce the number of questions in the entire survey, making it less burdensome for respondents to provide complete reports.

6. Future Work

Although the supposed neutrality of the survey as an impartial data collection tool means that all respondents have the same (or very similar) survey experiences, this rigid structure can also hinder the natural flow of information that occurs in a conversation (Suchman and Jordan 1990). Often for a participant, a particular event influences their answers for multiple questions. However, unless a direct question about this event appears in the survey, they have to answer many repetitive questions that could have been avoided in a conversation. Learning a latent structure of participants' answers in a survey could be a

step toward uncovering these hidden events that determine the answers to multiple questions, and DQO could use this knowledge to guide question selection as well.

As we mentioned at the outset, the cognitive aspects of survey methodology movement that originated in the 1980s (Jabine et al. 1984; Tanur 1992) raised issues with the traditional approach to survey questionnaire design, which keeps order fixed for all respondents and which measures the same quantities at different points in time. The need to reduce respondent burden and to keep respondents engaged in online surveys is raising a complementary set of issues that are now being addressed under the rubric of adaptive survey design. These two perspectives do need to be reconciled in some fashion.

In this article, we considered the prediction-focused implementation of DQO as a special case of the more general survey-taking setting. However, given typical survey respondents' disengagement from surveys and declining survey response rates, maybe a new paradigm of survey collection, in which respondents get something useful to them out of answering a survey, could motivate participants to provide complete and accurate responses (e.g., (Marcus et al. 2007; Angelovska and Mavrikiou 2013)). However, one clear downside to this approach is that giving respondents information that comes from the survey they are currently answering contaminates their response. For example, suppose that a person is answering questions about their energy-using habits to receive a personalized energy estimate, as in Subsection 3.1. Their current estimate for natural gas consumption is higher than they would like, and the next question asks for their preferred temperature in the winter. Because they do not want their estimate for natural gas usage to climb even higher, the respondent gives an optimistically-low value for preferred temperature. The uncertainty associated with these predictions can also influence a user's decision to continue answering questions: once a participant feels that their given prediction is certain enough, they may stop answering questions. Depending on the purpose of the survey (namely, whether its chief goal is to provide information to or to collect information from the respondent), this type of breakoff may or may not be bad.

7. Conclusion

Dynamic question ordering – that is, choosing which question to ask a survey respondent next, depending on their answers to previous questions – can improve survey quality in two key ways. First, giving participants personalized question orders can engage them and motivate them to complete the survey. Second, eliciting the most relevant information for a particular respondent upfront can improve the quality of imputations for unanswered questions if the respondent breaks off before completing the questionnaire. For some surveys, the goal is only to estimate a value for each respondent. In this case, it is not even necessary for the participant to answer all questions – it is sufficient for them to answer a subset that will ensure a confident prediction.

We present a general framework for dynamic question ordering in online surveys that sequentially considers which question to ask a respondent next, based on their previous answers, trading off the expected utility of having an answer to that question with the cost of asking that question. The definition of "utility" for an answer depends on the survey and its purpose. Examples include information gain, response probability, (negative) breakoff probability, or certainty of the subsequent prediction. Similarly, the definition of question

"cost" also depends on the survey. Examples include difficulty to the user of answering the question, (negative) likelihood of answering (since respondents may be reluctant to respond to sensitive questions, even if they are easy to answer), or breakoff rates of individual questions.

We illustrated two examples of this DQO framework. The first was for a prediction-oriented survey – providing prospective tenants with personalized energy estimates in potential homes. In this application, we found that asking users, on average, 21% of 30 questions could provide certain and accurate predictions at only 26% of the cost of the full-feature model, and that there was no fixed order of questions that was optimal across all users. Our second case study focused on traditional survey data collection on the Survey of Income and Program Participation. We selected subsequent questions that maximized conditional entropy to get a "representative" set of answers from participants early on. We simulated survey breakoff and found that, compared to the standard SIPP question order, the dynamic order delays breakoff, better recovers values for unanswered items that must be imputed, and achieves better quality survey estimates (lower bias and variance). Then, we discussed ways that dynamic question ordering could improve quality in computerized national surveys, focusing on unique aspects of each survey that DQO must take into account.

As more surveys move online or to computerized modes, dynamic question ordering can improve survey results at scale and at low cost to the data collectors. DQO trades off the utility from having an answer to a question with its cost and sequentially requests feature values in order to make useful, confident predictions and gather survey data with the resources users are willing and able to provide.

8. References

Achtyes, E.D., S. Halstead, L. Smart, T. Moore, E. Frank, D.J. Kupfer, and R. Gibbons. 2015. "Validation of Computerized Adaptive Testing in an Outpatient Nonacademic Setting: The VOCATIONS Trial." *Psychiatric Services* 66(10): 1091–1096. Doi: http://dx.doi.org/10.1176/appi.ps.201400390.

Almirall, D., S.N. Compton, M. Gunlicks-Stoessel, N. Duan, and S.A. Murphy. 2012. "Designing a Pilot Sequential Multiple Assignment Randomized Trial for Developing an Adaptive Treatment Strategy." *Statistics in Medicine* 31(17): 1887–1902. Doi: http://dx.doi.org/10.1002/sim.4512.

Angelovska, J. and P.M. Mavrikiou. 2013. *Can Creative Web Survey Questionnaire Design Improve the Response Quality?* University of Amsterdam, AIAS Working Paper, 131. Available at: http://archive.uva-aias.net/uploadedfiles/publications/AIASWP131-1.pdf (accessed April 14, 2017).

Barge, S. and H. Gehlbach. 2012. "Using the Theory of Satisficing to Evaluate the Quality of Survey Data." *Research in Higher Education* 53(2): 182–200. Doi: http://dx.doi.org/10.1007/s11162-011-9251-2.

Benedetto, G., M. Stinson, and J.M. Abowd. 2013. *The Creation and Use of the SIPP Synthetic Beta. U.S. Census Bureau*. Available at: http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribenontechnical.pdf (accessed April 14, 2017).

Bouamrane, M.-M., A. Rector, and M. Hurrell. 2008. *Gathering Precise Patient Medical History with an Ontology-Driven Adaptive Questionnaire*. In Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems, 539–541, 17–19 June 2008, Jyväskylä, Finland. Doi: http://dx.doi.org/10.1109/CBMS.2008.24.

Bradburn, N. 1978. "Respondent Burden." In Proceedings of the American Statistical Association, Survey Research Methods Section, 35–40. Available at: http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1978_007.pdf (accessed April 14, 2017).

Bureau of Justice Statistics. 2014. *National Crime and Victimization Survey: Technical Documentation*. Available at: http://www.bjs.gov/content/pub/pdf/ncvstd13.pdf (accessed April 14, 2017).

Clark, S.L. 2014. *American Community Survey Item Nonresponse Rates: Mail versus internet*. American Community Survey Research and Evaluation Program (March). Available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/acs/2014Clark01.pdf (accessed April 14, 2017).

Cohn, D.A., Z. Ghahramani, and M.I. Jordan. 1996. "Active Learning with Statistical Models." *Journal of Artificial Intelligence Research* 4: 129–145.

Collins, L.M., S.A. Murphy, V.N. Nair, N. Vijay, and V.J. Strecher. 2005. "A Strategy for Optimizing and Evaluating Behavioral Interventions." *Annals of Behavioral Medicine* 30(1): 65–73. Doi: http://dx.doi.org/10.1207/s15324796abm3001_8.

Collins, L.M., S.A. Murphy, and V. Strecher. 2007. "The Multiphase Optimization Strategy (MOST) and the Sequential Multiple Assignment Randomized Trial (SMART): New Methods for More Potent eHealth Interventions." *American Journal of Preventive Medicine* 32(5): S112–S118. Doi: http://dx.doi.org/10.1016/j.amepre.2007.01.022.

Cook, C., F. Heath, and R.L. Thompson. 2000. "A Meta-Analysis of Response Rates in Web- or Internet-Based Surveys." *Educational and Psychological Measurement* 60(6): 821–836. Doi: http://dx.doi.org/10.1177/00131640021970934.

Couper, M.P., G.L. Alexander, N. Zhang, R.J.A. Little, N. Maddy, M.A. Nowak, J.B. McClure, J.J. Calvi, S.J. Rolnick, M.A. Stopponi, and C.C. Johnson. 2010. "Engagement and Retention: Measuring Breadth and Depth of Participant Use of an Online Intervention." *Journal of Medical Internet Research* 12(4): e52. Doi: http://dx.doi.org/10.2196/jmir.1430.

Cover, T.M. and P.E. Hart. 1967. "Nearest Neighbor Pattern Classification." *IEEE Transactions on Information Theory* 13(1): 21–27. Doi: http://dx.doi.org/10.1109/tit.1967.1053964.

Cover, T.M. and J.A. Thomas. 2012. *Elements of Information Theory*. John Wiley & Sons. Doi: http://dx.doi.org/10.1002/047174882x.

Dietz, T., G.T. Gardner, J. Gilligan, P.C. Stern, and M.P. Vandenbergh. 2009. "Household Actions Can Provide a Behavioral Wedge to Rapidly Reduce U.S. Carbon Emissions." In Proceedings of the National Academy of Sciences, November 3, 2009. 106: 18452–18456. National Academy of Sciences. Doi: http://dx.doi.org/10.1073/pnas.0908738106.

Early, K., S. Fienberg, and J. Mankoff. 2016. "Test-Time Feature Ordering with FOCUS: Interactive Predictions with Minimal User Burden." In Proceedings of the 2016 ACM

International Joint Conference on Pervasive and Ubiquitous Computing, September 12–16, Heidelberg, Germany, 992–1003. Doi: http://dx.doi.org/10.1145/2971648.2971748.

Ferrucci, D., A. Levas, S. Bagchi, D. Gondek, and E.T. Mueller. 2013. "Watson: Beyond Jeopardy!" *Artificial Intelligence* 199: 93–105. Doi: http://dx.doi.org/10.1016/j.artint.2012.06.009.

Fliege, H., J. Becker, O.B. Walter, J.B. Bjorner, B.F. Klapp, and M. Rose. 2005. "Development of a Computer-Adaptive Test for Depression (D-CAT)." *Quality of Life Research* 14(10): 2277–2291. Doi: http://dx.doi.org/10.1007/s11136-005-6651-9.

Fricker, S., T. Yan, and S. Tsai. 2014. "Response Burden: What Predicts it and who is Burdened Out." In Proceedings of the American Association for Public Opinion Research, May 15–18, 2014, Anaheim, California. 4568–4577. Available at: http://ww2.amstat.org/sections/SRMS/Proceedings/y2014/Files/400298500838.pdf (accessed April 14, 2017).

Fuller, W.A. 2009. *Measurement Error Models*. New York: Wiley. Doi: http://dx.doi.org/10.1002/9780470316665.

Galesic, M. 2006. "Dropouts on the Eeb: Effects of Interest and Burden Experienced During an Online Survey." *Journal of Official Statistics* 22(2): 313–328.

Galesic, M. and M. Bosnjak. 2009. "Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey." *Public Opinion Quarterly* 73(2): 349–360. Doi: http://dx.doi.org/10.1093/poq/nfp031.

Gardner, W., K. Shear, K.J. Kelleher, K.A. Pajer, O. Mammen, D. Buysse, and E. Frank. 2004. "Computerized Adaptive Measurement of Depression: A Simulation Study." *BMC Psychiatry* 4(1): 13–23. Doi: http://dx.doi.org/10.1186/1471-244x-4-13.

Gibbons, R.D., D.J. Weiss, E. Frank, and D.J. Kupfer. 2016. "Computerized Adaptive Diagnosis and Testing of Mental Health Disorders." *Annual Review of Clinical Psychology* 12(1): 83–104. Doi: http://dx.doi.org/10.1146/annurev-clinpsy-021815-093634.

Gibbons, R.D., D.J. Weiss, D.J. Kupfer, E. Frank, A. Fagiolini, V.J. Grochocinski, and J.C. Immekus. 2008. "Using Computerized Adaptive Testing to Reduce the Burden of Mental Health Assessment." *Psychiatric Services* 59(4): 361–368. Doi: http://dx.doi.org/10.1176/ps.2008.59.4.361.

Gibbons, R.D., D.J. Weiss, P.A. Pilkonis, E. Frank, T. Moore, J.B. Kim, and D.J. Kupfer. 2012. "Development of a Computerized Adaptive Test for Depression." *Archives of General Psychiatry* 69(11): 1104–1112. Doi: http://dx.doi.org/10.1001/archgenpsychiatry.2012.14.

Griffin, D. and D. Nelson. 2014. "Reducing Respondent Burden in the ACS's Computer Assisted Personal Visit Interviewing Operation - Phase 1 Results." Available at: https://www.census.gov/content/dam/Census/library/working-papers/2014/acs/2014_Griffin_02.pdf (accessed April 14, 2017).

Groves, R.M. and S.G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society:* Series A (Statistics in Society) 169(3): 439–457. Doi: http://dx.doi.org/10.1111/j.1467-985x.2006.00423.x.

Harrell, F.E. 2001. *Regression Modeling Strategies*. New York: Springer. Doi: http://dx.doi.org/10.1007/978-1-4757-3462-1.

He, H., H. III Daume, and J. Eisner. 2012. "Cost-Sensitive Dynamic Feature Selection." In Inferning 2012: ICML workshop on interaction between inference and learning. Edinburgh, Scotland. Available at: https://www.cs.jhu.edu/~jason/papers/he+al.icmlw12.pdf (accessed April 14, 2017).

Horwitz, R., J. Tancreto, M.F. Zelenak, and M. Davis. 2012. "Data Quality Assessment of the American Community Survey Internet Response Data." Available at: https://www.census.gov/content/dam/Census/library/working-papers/2012/acs/2012.Horwitz.02.pdf (accessed April 14, 2017).

Jabine, T.B., M.L. Straf, J.M. Tanur, and R. Tourangeau. 1984. *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington, DC: National Academies Press. Doi: http://dx.doi.org/10.2307/2289187.

Kaczmirek, L. 2008. "Human-Survey Interaction: Usability and Nonresponse in Online Surveys" (Doctoral dissertation, Universität Mannheim). Available at: http://ub-madoc.bib.uni-mannheim.de/2150/1/kaczmirek2008.pdf (accessed April 14, 2017).

Kamakura, W.A. and S.K. Balasubramanian. 1989. "Tailored Interviewing: An Application of Item Response Theory for Personality Measurement." *Journal of Personality Assessment* 53(3): 502–519. Doi: http://dx.doi.org/10.1207/s15327752jpa5303.8.

Kapelner, A. and D. Chandler. 2010. "Preventing Satisficing in Online Surveys." In Proceedings of CrowdConf. San Francsico, California. Available at: http://s3.amazonaws.com/academia.edu.documents/30740949/kapcha.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1499626168&Signature=0zH5TV1LDugz3h9FlSRJy5lRp38%3D&response-content-disposition=inline%3B%20filename%3DPreventing_Satisficing_in_Online_Surveys.pdf (accessed July 9, 2017).

Karayev, S., T. Baumgartner, M. Fritz, and T. Darrell. 2012. "Timely Object Recognition." In *Advances in Neural Information Processing Systems* 25: 890–898. Lake Tahoe, Nevada. Available at: https://papers.nips.cc/paper/4712-timely-object-recognition.pdf (accessed April 14, 2017).

Kilbourne, A.M., D. Almirall, D. Eisenberg, J. Waxmonsky, D.E. Goodrich, J.C. Fortney, J.E. Kirchner, L.I. Solberg, D. Main, M.S. Bauer, J. Kyle, S.A. Murphy, K.M. Nord, and M.R. Thomas. 2014. "Protocol: Adaptive Implementation of Effective Programs Trial (ADEPT): Cluster Randomized SMART Trial Comparing a Standard Versus Enhanced Implementation Strategy to Improve Outcomes of a Mood Disorders Program." *Implementation Science* 9(132): 1–14. Doi: http://dx.doi.org/10.1186/s13012-014-0132-x.

Laurie, H., and L. Scott. 1999. "Strategies for Reducing Nonresponse in a Longitudinal Panel Survey." *Journal of Official Statistics* 15(2): 269–282.

Lavori, P.W., R. Dawson, and A.J. Rush. 2000. "Flexible Treatment Strategies in Chronic Disease: Clinical and Research Implications." *Biological Psychiatry* 48(6): 605–614. Doi: http://dx.doi.org/10.1016/s0006-3223(00)00946-x.

Lord, F.M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates. Doi: http://dx.doi.org/10.4324/9780203056615.

Manfreda, K.L. and V. Vehovar. 2002. "Survey Design Features Influencing Response Rates in Web Surveys." In The International Conference on Improving Surveys. Copenhagen, Denmark. Available at: http://www.websm.org/uploadi/editor/Lozar_Vehovar_2001_Survey_design.pdf (accessed April 14, 2017).

Marcus, B., M. Bosnjak, S. Lindner, S. Pilischenko, and A. Schütz. 2007. "Compensating for Low Topic Interest and Long Surveys a Field Experiment on Nonresponse in Web Surveys." *Social Science Computer Review* 25(3): 372–383. Doi: http://dx.doi.org/10.1177/0894439307297606.

McFarland, S.G. 1981. "Effects of Question Order on Survey Responses." *Public Opinion Quarterly* 45(2): 208–215. Doi: http://dx.doi.org/10.1086/268651.

Montgomery, J.M., and J. Cutler. 2013. "Computerized Adaptive Testing for Public Opinion Surveys." *Political Analysis* 21(2): 172–192. Doi: http://dx.doi.org/10.1093/pan/mps060.

Murphy, S.A. 2005. "An Experimental Design for the Development of Adaptive Treatment Strategies." *Statistics in Medicine* 24: 1455–1481. Doi: http://dx.doi.org/10.1002/sim.2022.

National Center for Health Statistics. 2014. "Survey Description: National Health Interview Survey." Available at: http://ftp.cdc.gov/pub/Health_Statistics/NCHs/Dataset_Documentation/NHIS/2014/srvydesc.pdf (accessed April 14, 2017).

Pandey, D., M. Agrawal, and J.S. Pandey. 2011. "Carbon Footprint: Current Methods of Estimation." *Environmental Monitoring and Assessment* 178(1–4): 135–160. Doi: http://dx.doi.org/10.1007/s10661-010-1678-y.

Peytchev, A. 2009. "Survey Breakoff." *Public Opinion Quarterly* 73(1): 74–97. Doi: http://dx.doi.org/10.1093/ poq/nfp014.

Pitkow, J.E. and M.M. Recker. 1995. "Using the Web as a Survey Tool: Results from the Second WWW User Survey." *Computer Networks and ISDN Systems* 27(6): 809–822.

Porter, S.R. 2004. "Raising Response Rates: What Works?" *New Directions for Institutional Research* 2004(121): 5–21. Doi: http://dx.doi.org/10.1002/ir.97.

Rubin, D.B. 2004. *Multiple Imputation for Nonresponse in Surveys* (Vol. 81). John Wiley & Sons. Doi: http://dx.doi.org/10.1002/9780470316696.

Rush, A.J., M. Fava, S.R. Wisniewski, P.W. Lavori, M.H. Trivedi, H.A. Sackeim, M.E. Thase, A.A. Nierenberg, F.M. Quitkin, T.M. Kashner, and D.J. Kupfher. 2004. "Sequenced Treatment Alternatives to Relieve Depression (STAR* D): Rationale and Design." *Contemporary Clinical Trials* 25(1): 119–142. Doi: http://dx.doi.org/10.1016/S0197-2456(03)00112-0.

Schouten, B., M. Calinescu, and A. Luiten. 2013. "Optimizing Quality of Response Through Adaptive Survey Designs." *Survey Methodology* 39(1): 29–58.

Shih, T.-H., and X. Fan. 2008. "Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis." *Field Methods* 20(3): 249–271. Doi: http://dx.doi.org/10.1177/1525822x08317085.

Singh, J., R.D. Howell, and G.K. Rhoads. 1990. "Adaptive Designs for Likert-Type Data: An Approach for Implementing Marketing Surveys." *Journal of Marketing Research* 27(3): 304–321. Doi: http://dx.doi.org/10.2307/3172588.

Stroup, T.S., J.P. McEvoy, M.S. Swartz, M.J. Byerly, I.D. Glick, J.M. Canive, and . . . J.A. Lieberman. 2003. "The National Institute of Mental Health Clinical Antipsychotic

Trials of Intervention Effectiveness (CATIE) Project: Schizophrenia Trial Design and Protocol Development." *Schizophrenia Bulletin* 29(1): 15–31. Doi: http://dx.doi.org/10.1093/oxfordjournals.schbul.a006986.

Suchman, L. and B. Jordan. 1990. "Interactional Troubles in Face-to-Face Survey Interviews (with discussion)." *Journal of the American Statistical Association* 85(409): 232–253. Doi: http://dx.doi.org/10.1080/01621459.1990.10475331.

Sudman, S., N.M. Bradburn, and N. Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Sun, M., F. Li, J. Lee, K. Zhou, G. Lebanon, and H. Zha. 2013. "Learning Multiple-Question Decision Trees for Cold-Start Recommendation." In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, 445–454. Rome, Italy: ACM. Doi: http://dx.doi.org/10.1145/2433396.2433451.

Tanur, J.M. 1992. *Questions About Questions: Inquiries into the Cognitive Bases of Surveys*. New York: Russell Sage Foundation. Doi: http://dx.doi.org/10.2307/2075046.

Tourangeau, R. 1984. "Cognitive Sciences and Survey Methods." In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, edited by T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau, 73–100. Washington, DC: National Academies Press.

Tropp, J.A. 2004. "Greed is Good: Algorithmic Results for Sparse Approximation." *IEEE Transactions on Information Theory* 50(10): 2231–2242. Doi: http://dx.doi.org/10.1109/tit.2004.834793.

U.S. Bureau of the Census. 2006. *Design and methodology: Current Population Survey*. Available at: http://www.nber.org/cps/tp-66.pdf (accessed April 14, 2017).

U.S. Bureau of the Census. 2013. *American Housing Survey for the United States*. Washington, D.C: U.S. Government Printing Office.

U.S. Bureau of the Census. 2014a. *American Community Survey: Design and methodology*. Available at: http://www2.census.gov/programs-surveys/acs/methodology/designandmethodology/acs_design_methodology_report_2014.pdf (accessed April 14, 2017).

U.S. Bureau of the Census. 2014b. *Survey of Income and Program Participation*. Available at: www.census.gov/programs-surveys/sipp (accessed April 14, 2017).

U.S. Bureau of the Census. 2016. *American Community Survey*. Available at: https://www.census.gov/programs-surveys/acs/ (accessed April 14, 2017).

U.S. Energy Information Administration. 2009. *Residential Energy Consumption Survey 2009*. Available at: www.eia.gov/consumption/residential/data/2009/. (accessed April 14, 2017).

Walston, J.T., R.W. Lissitz, and L.M. Rudner. 2006. "The Influence of Web-Based Questionnaire Presentation Variations on Survey Cooperation and Perceptions of Survey Quality." *Journal of Official Statistics* 22(2): 271–291.

Wang, L., A. Rotnitzky, X. Lin, R.E. Millikan, and P.F. Thall. 2012. "Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer." *Journal of the American Statistical Association* 107(498): 493–508. Doi: http://dx.doi.org/10.1080/01621459.2011.641416.

Weisberg, S. 2014. *Applied Linear Regression* (4th ed.). New York: Wiley. Doi: http://dx.doi.org/10.1002/0471704091.

Weiss, D.J. 1982. "Improving Measurement Quality and Efficiency with Adaptive Testing." *Applied Psychological Measurement* 6(4): 473–492. Doi: http://dx.doi.org/10.1177/014662168200600408.

Weiss, D.J. and G.G. Kingsbury. 1984. "Application of Computerized Adaptive Testing to Educational Eroblems." *Journal of Educational Measurement* 21(4): 361–375. Doi: http://dx.doi.org/10.1111/j.1745-3984.1984.tb01040.x.

Yu, E.C., S. Fricker, and B. Kopp. 2015. "Can Survey Instructions Relieve Respondent Burden?" In AAPOR. Hollywood, Florida. Available at: http://www.bls.gov/osmr/pdf/st150260.pdf (accessed April 14, 2017).

Received March 2016
Revised April 2017
Accepted May 2017